Rigorous Validation of Systems Security Engineering Analytics

Thomas Llansó
Johns Hopkins University APL
thomas.llanso@jhuapl.edu

Martha McNeil
Johns Hopkins University APL
martha.mcneil@jhuapl.edu

Jessie Jamieson
Johns Hopkins University APL
jessie.jamieson@jhuapl.edu

Abstract

In response to the asymmetric advantage that attackers enjoy over defenders in cyber systems, the cyber community has generated a steady stream of cybersecurity-related frameworks, methodologies, analytics, and "best practices" lists. However, these artifacts almost never undergo rigorous validation of their efficacy but instead tend to be accepted on faith, to, we suggest, our collective detriment based on evidence of continued attacker success. But what would rigorous validation look like, and can we afford it? This paper describes the design and estimates the cost of a controlled experiment whose goal is to determine the effectiveness of an exemplar systems security analytic. Given the significant footprint that humans play in cyber systems (e.g., their design, use, attack, and defense), any such experiment must necessarily take into account and control for variable human behavior. Thus, the paper reinforces the argument that cybersecurity can be understood as a hybrid discipline with strong technical and human dimensions.

1. Introduction

Similar to other engineering disciplines, cybersecurity is a highly technical field that employs a broad range of mechanisms, such as encryption, authentication, access control, intrusion detection, and firewalls. Despite the technical nature of the field, the influence of humans on cybersecurity also looms large, as people design, use, attack, and defend cyber-enabled systems. Thus, while not always appreciated as such, cybersecurity is also a sociotechnical endeavor.

An illustration of the human dimension of cyber is found in cybersecurity architecture and design. To create robust and secure cyber architectures, systems security engineers (SSEs) must determine the best way to *combine* technical mechanisms, such as those listed above, into cohesive and affordable architectures that are responsive to anticipated threat actions. Historically, SSEs have been left do what "feels right," with decisions driven more by intuition and heuristics than by strong theoretical underpinnings supported by a body of empirical evidence [1][2]. In recent years, SSEs have increasingly sought out estimates of business/mission risk and resilience to inform their engineering decisions as well threat-based approaches for selecting targeted defensive mechanisms [3][4]. However, developing the estimates and making decisions have typically required tedious and subjective manual scoring and ranking processes. Unfortunately, research suggests uncomfortably high scoring variance exists across teams that conduct such analyses [5]–[7].

Consequently, interest has grown in the use of automated decision support analytics to help SSEs to more reliably and systematically specify architectures. The question naturally arises as to the effectiveness of these analytics: are we better off using them than not? And if so, by how much? After all, while these analytics might help us to be more consistent in our choices, they might also help us to be consistently *wrong*; that is, they might help improve precision but not accuracy [8]. We therefore have a validation concern.

Validation issues have been in the news in recent years in, for example, the social sciences context, with many examples of social science research having proven difficult to replicate [9][10]. How can we validate candidate architecture-level cybersecurity analytics in a way that has a chance of successful replication by others? This paper discusses validation for an exemplar cybersecurity analytic as an example of the larger validation concern in cybersecurity. By cybersecurity analytic we mean an automated procedure to estimate some quantity of interest to SSEs, such as risk or resilience estimates. The contribution of the paper is the description of a rigorous experiment for analytic validation that explicitly considers control for human variance. The paper also estimates the cost and potential benefits of running such experiments.

In the language of clinical trials, the experiment we describe is a randomized, double-blind design whose goal is to establish the safety and efficacy of a cybersecurity treatment. Here, the treatment is use of a cyber analytic for decision support purposes during architecture development. By safety, we mean that the

URI: https://hdl.handle.net/10125/80252
978-0-9981331-5-7

treatment does not unwittingly make a system *less* secure, as can happen, for example, if constituent cybersecurity mechanisms are themselves vulnerable and/or are combined in insecure ways. If the experiment is conducted with sufficient statistical power and in a manner that controls for variability, especially with respect to human participants, then one can begin to make more confident, even causal, claims of effectiveness that others, in the tradition of science, can then attempt to replicate. Such "gold standard" experiments are among the most expensive forms of validation, and they are rare in cybersecurity [11] even as they are common in other scientific fields, including the social sciences. Part of the reason they are uncommon in cybersecurity, we suggest, is that the science of cybersecurity has been outrun by the furious rush to deploy systems over the past several decades, with the rapid and mass adoption of the Internet, the World Wide Web, and "Internet of Things." The urgency in cybersecurity to "do something now" has, in our estimation, created a cultural bias against running costly experiments that may result in negative findings.

The rest of the paper is organized as follows. We begin by discussing definitions and related work. Next, to illustrate the experimental design for analytic validation, we describe an illustrative analytic as the target of the experiment. Subsequent sections cover research method selection, experimental design, threats to validity, controls for those threats, and factorial design considerations. Before concluding, we discuss cost versus benefit.

2. Definitions

Validation. The term validation has multiple definitions in the literature. We adopt the International Standards Organization definition for validation [12]: "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled." Validation considers whether a system addresses the intended mission/business needs. A closely related term, verification, is the "confirmation, through the provision of objective evidence, that the specified requirements have been fulfilled" [12], i.e. that the system is built to match its specifications [13]. Barry Boehm [13] famously summed up the difference between validation and verification as a determination of whether one has built the right system (to satisfy a user need) versus having built the system right (as specified).

Validity. Despite being derived from the same root word, validity and validation are different, albeit related. We adopt Trochim's definition of validity [14] "the best available approximation to the truth of a given proposition, inference, or conclusion." Trochim describes many types of validity (e.g., face, construct, internal, external, sampling). Researchers must design studies with validity in mind, considering sampling techniques, confounding factors, and other elements that may affect research validity.

Repeatability and Reproducibility. Repeatability and reproducibility are "cornerstones of the scientific process, necessary for avoiding dissemination of flawed results" [15]. Repeatability is the ability to run the same experiment using the same method and obtain the same result. It is usually done by the same researcher and is used to ensure that the research foundation is stable. Reproducibility is the ability by independent researchers to confirm the experimental results. The latter is more powerful as successful reproduction can support generalizability of the research, whereas failed reproduction can expose errors [16].

3. Related Work

This section surveys work relevant to validation in cybersecurity, including validation methods, challenges to science in cybersecurity, examples of progress towards achieving scientific rigor in cybersecurity, repeatability and reproducibility challenges, and the social dimension of cybersecurity.

Validation methods. Validation is not a monolithic pass/fail endeavor that occurs just prior to use. Instead, we take steps towards validation across the system lifecycle. For example, we use peer review, a form of validation that relies mainly on face validity, to assure ourselves that we are building *the right system* during requirements, analysis, design, and implementation. Likewise, pilot deployment (i.e. "taking it out for a spin") can bolster validation efforts, especially when performed with the rigor of methods such as action research or case study. While they provide useful insights, such studies provide only point-in-time information and are usually weak in external validity/generalizability.

One can also gain insights when system usage is combined with survey research supported by formal models, such as the Technology Acceptance Model (TAM) [17], [18] and DeLone and McLean's Information System (IS) Success Model (D&M) [19], [20]. TAM allows inferences between users' perceived usefulness/ease of use of an IS and their intention/actual usage. By contrast, D&M permits drawing inferences between user satisfaction/actual usage and individual/organizational impact. But what is lacking so far in these techniques is causality, i.e. finding support that the system under evaluation has the desired impact on the problem it is intended to solve. Causality is the strongest indicator that we have built *the right system*.

For this paper, that system is an analytic intended to support architectural cybersecurity decisions.

Challenges to Science in Cybersecurity. Starting in the 2000s, a significant uptick occurred in promoting the application of scientific principles to cybersecurity as a way to advance the state of the art [1][11], [21]–[28]. Early on, one author noted that "In order to claim scientifically valid and justifiable results, computer security experiments must follow the scientific method" and only then can "those who make decisions about the security of electronic voting machines, hospital operating room equipment, and airplane software" [27] trust in the technology. In fact, advocacy for experimentation in the larger realm of computer science dates back to at least 1998 [29] when Tichy promoted the practice and rebutted eight fallacious arguments against doing so. More recently, Longstaff et al. asserted that "Understanding the underlying causality of information security could allow [the field] to leapfrog ahead in our solutions rather than just improve incrementally in a never-ending arms race." [11] Unfortunately, the authors also noted that while "finding agreement in the use of the scientific method is practically universal, finding participation in the scientific method is rare" [11]. In 2017, a researcher decried a tendency to rely on the "uniqueness" of security as an excuse for avoiding "practices ubiquitous elsewhere in science" [2]. Still, Landwehr [21] asserts that it is not too late for cybersecurity to embrace science, noting that "scientific foundations frequently follow, rather than precede, the development of practical, deployable solutions to particular problems".

Scientific Experimentation in Cybersecurity. Peisert and Bishop [27] and Carroll et al. [28] describe key elements of the rigor necessary in scientific cybersecurity experimentation, including falsifiable hypotheses, reproducible results, controlled independent and dependent variables, and accounting for confounds [27], [28]. Coopamootoo, et al. present a "design and reporting" toolkit that has nine key indicators of rigorous design for cybersecurity experiments [30]. Although the application of science in cybersecurity may not have advanced as far as we might have wished in the preceding decades (see [11], [23], [28], [29], [31] for some possible reasons why), some noteworthy experimental studies are in the literature [5][32]–[35].

Repeatability and Reproducibility Challenges. While repeatability and reproducibility are fundamental to achieving rigor in cyber research, the information needed by other researchers is rarely made available when such research is published [16]. Further, Carroll et al. note that "Reproducibility is a challenge in cyber security due to the highly complex interdependent systems" studied where "the state space is massive" [28]. As mentioned, human variability confounds repeatability, yet, as Tichy points out, other science-based fields, notably medicine and psychology, have found ways to cope [29].

Cybersecurity and Social Science. Maxion et al. suggest that studying cybersecurity only through the lenses of engineering and the physical sciences may be the wrong model for many research questions. Better models may exist in "the social sciences which deal with the same sorts of complex and uncontrollable environments as cyber security, but nonetheless have well established methodologies for performing experiments, analyzing data and reporting results so that others may build upon them" [31]. A 2017 National Academy of Sciences report concurs and encourages cross-disciplinary research collaborations between cybersecurity researchers and researchers in domains including "economics (e.g., incentives, resources), sociology (e.g., social networks, norms, criminology), psychology (e.g., motivation, perception, user interfaces)," and more [36]. Passwords and security policies are intensely human aspects of cybersecurity; hence, it is not surprising to find examples of research blending cyber with the social sciences, a few of which include [37]–[39].

4. Analytic To Be Validated

To provide a concrete example for our proposed experiment, we selected an exemplar analytic tool called BluGen [40] as the validation target. We chose BluGen because it is published in the open literature and, as its authors, we are familiar with its details. However, the approach applies to any similar analytic provided there exists a clear way of measuring outcomes based on analytic use. Briefly, BluGen estimates mission/business risk due to cyber effects and then recommends mitigations to help lower risk based on user-specified risk tolerance threshold levels.

System security engineers may employ such an analytic to inform their decisions based on risk and mitigation recommendations, leading to an engineered system (Figure 1). The figure references "Blue Team-E," which we define as the team of security engineers responsible for selecting and engineering in the mitigations to a system. Later we reference "Blue Team-D," defined as the team that monitors and actively defends the deployed system using the engineered defenses and other tools.

Figure 1: Analytic Informing Engineering Decisions

5. Measuring Outcomes

The experiment's goal is to find support for a causal relationship between use of the cybersecurity analytic, the independent variable, and improved mission/business outcomes, the dependent variable. Organizations employ cyber-enabled systems to achieve mission/business goals; thus, cybersecurity is primarily focused on maximizing "mission" success despite adverse cyber events, such as malicious attacks.

Put another way, we are trying to minimize the number of consequential attacks, CA. The CA count is relative to a given target cyber system, s, over a defined operational time interval [t1, t2]. During the interval, cyber attacks may occur, some of which are 'consequential.' By consequential we mean an attack in which the performance for one or more defined mission essential functions of s drop below their associated minimum threshold values at one or more points over [t1, t2]. The proposed experiment evaluates the hypothesis, H_A, and associated null hypothesis, H_0 for the exemplar analytic:

H_0: *Use of the analytic does not decrease CA.*
H_A: *Use of the analytic decreases CA.*

6. Research Method Selection

Consistent with Herbert Simon's "Sciences of the Artificial," [41] cybersecurity analytics are constructed artifacts and should be evaluated as such. Design Science Research (e.g., [42][43]) is helpful in this regard. Figure 2 from Venable, et al. [44] identifies validation approaches for artifacts prior to (Ex Ante) and subsequent to (Ex Post) characterization.

DSR Evaluation Method Selection Framework	Ex Ante	Ex Post
Naturalistic	•Action Research •Focus Group	•Action Research •Case Study •Focus Group •Participant Observation •Ethnography •Phenomenology •Survey (qualitative or quantitative)
Artificial	•Mathematical or Logical Proof •Criteria-Based Evaluation •Lab Experiment •Computer Simulation	•Mathematical or Logical Proof •Lab Experiment •Role Playing Simulation •Computer Simulation •Field Experiment

Figure 2: Evaluation Framework [44]

Given that our analytic is human-designed and already exists, our focus is on methods found at the intersection of the "Artificial" row and "Ex Post" column of Figure 2. With causality in mind, we choose the "Lab Experiment" method from Figure 2.

7. Experimental Design

A posttest-only randomized experiment [14] is appropriate in our context (Figure 3). This design is useful for evaluating the internal validity of postulated cause-effect relationships.

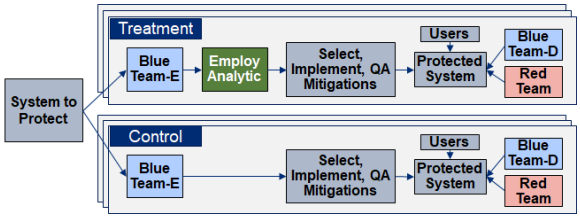

Figure 3: Experimental Design

In this context, the cause is a blue team's use of an analytic to inform the team's engineering decisions and the effect is a system whose overall protections are hypothesized to show a decrease (improvement) in CA over a system whose protections are derived and implemented manually. That is, the CA count is the posttest. Researchers randomly assign teams to treatment and control groups. Blue Team-E members obviously know whether or not they are assigned to treatment or control lines based on the use/non-use of the analytic in question; however, they are blind to the users, Blue Team-D, and the Red Teams involved with the system they engineer. Per the usual definition, the red teams mimic the anticipated adversary by researching and attempting to execute example attacks. Similarly, members of the Red Teams and Blue Team-D are blind to whether they are in the treatment or control line.

Establishing causality requires conformance to three requirements: temporal precedence, covariation of cause and effect, and lack of plausible alternative explanations [14]. Temporal precedence stems from using the analytic prior to security engineering design decisions per Figure 1. The effect, the CA count, is measured later when the system is deployed. Establishing covariance is also straightforward because we assert a binary condition, use/non-use of the analytic, where use is hypothesized to cause a decrease in CA on average compared to non-use. We cover alternative explanations below.

To test the hypothesis, we run the treatment and control lines from Figure 3 multiple times holding the system and time interval constant in each run but varying the users and blue teams and red teams, such that any team is used only once and is randomly assigned. Next, we calculate the two CA means, $\bar{x}_{treatment}$ and $\bar{x}_{control}$, as shown below.

$\bar{x}_{treatment}$ = CA sample mean of treatment groups
$\bar{x}_{control}$ = CA sample mean of control groups

The question is whether $\bar{x}_{treatment}$ - $\bar{x}_{control}$ is statistically significant. Using a level of significance, $\alpha = 0.05$, for example, then if $p < 0.05$, then we reject H_0 and find support for H_A, that is, the tool brought about a measurable decline in CA count on average.

8. Threats to Validity

Multiple questions arise in our efforts to control for threats to the validity of the experiment. Even if we measure a statistically significant effect supporting H_A, is the effect due to use of the analytic or is something else at play? That is, is the effect is due to some other confounding variable or set of variables? Alternative explanations are threats to internal validity and, hence, the causality argument. In addition, do the results generalize beyond the study context? Threats to external validity could work against broader applicability of the analytic considered.

We consider threats to validity in six categories: system to protect, blue teams, red teams, users, reference datasets employed by the analytics, and the analytics themselves (Table 1). We assign each threat an ID for later reference. The threats can be part of an alternate explanation for the observed effect or can limit the external validity of the experiment.

Table 1: Sample Threats to Validity

System to Protect	
S-T1	System is not representative of real-world cyber systems of interest (e.g., in type, complexity)
S-T2	System description is not an accurate representation of the target system
Blue Teams (both -E and -D)	
B-T1	Blue team-E is overly constrained by time/budget constraints, limiting analysis and protections
B-T2	Blue Teams-E/D are not adequately informed of threats the system faces
B-T3	Blue teams-E/D are not representative of the kinds of teams typically encountered
B-T4	Blue team-E did a poor job implementing defenses recommended by the analytic
B-T5	Blue team-E was overly tolerant of risks and therefore did not implement all recommendations
Red Teams	
R-T1	Red team did not adequately represent the anticipated threat
R-T2	Red team was not given enough time to research and then attack the target system
System Users	
U-T1	Users are not representative of real system's users
Reference Datasets Used by the Analytics	
D-T1	Asset type taxonomy is insufficiently detailed, and thus misses certain threats mapped threats
D-T2	Threat capabilities represented do not faithfully represent the anticipated threat actor
D-T3	Threat capabilities are incorrectly mapped to applicable asset types
D-T4	Defensive capabilities are incorrectly mapped to threats they are intended to mitigate
D-T5	The set of defensive capabilities is incomplete compared to capabilities available more broadly.
Analytics	
A-T1	Analytic output is not useful and/or is ambiguous to blue teams

9. Controlling for Threats

Table 2 provides examples of possible controls for the threats described in Table 1.

Table 2: Controlling for Threats to Validity

Control	Threats
Adjust system types evaluated	S-T1
Apply multiple checks on system description (e.g., scan system if able, interview experts, review documentation)	S-T2
Establish a fixed, realistic time period and budget within which blue teams operate	B-T1
Ensure the blue teams are thoroughly briefed on the threat and that the team is confident in its ability to mimic the threat	B-T2
Employ blue teams with similar experience, team sizes, and composition	B-T3 B-T4 R-T1
Assign a consistent risk tolerance threshold for all blue teams to use.	B-T5
Allow a realistic and consistent time on target for the red teams	R-T2
Employ actual or representative users of the system during the experiment	U-T1
To help establish pedigree, match reference datasets to other sources and peer review mappings of threats↔asset types and mitigations↔threats	D-T1 D-T2 D-T3 D-T4 D-T5
Improve clarity of analytic descriptions and improve user training.	A-T1

10. Factorial Design Possibilities

A factorial design with blocking can help tease out the effects of individual control variables on CA. An example appears in Table 3. A full factorial experiment has 64 unique combinations of the six variables from Table 3. The expense of repeating the experiment for this many combinations would be high. However, holding certain factors at a fixed level as shown in Table 4 results in a significant reduction of combinations from 64 to 4.

Execution of the experiment in certain domains might further alter the combinations. For example, if the interest is primarily on cyber-physical systems (e.g., a weapons system, a sensor, avionics) and the red

team is asked to mimic the highest tier threat [45], the number of combinations reduces to 1. Of course, such reductions limit the breadth of external validity.

Table 3: Factorial Design

#	Factor	Level	Value
1	System type	1	IT enterprise
		2	Cyber-physical
2	Average team experience level	1	< 5 years
		2	>= 5 years
3	Time Allowed - Blue	1	2 weeks
		2	3 weeks
4	Use of active blue defenders	1	No
		2	Yes
5	Time Allowed - Red	1	1 week
		2	2 weeks
6	Red Team Mimic	1	Mid-Tier
		2	High-Tier

Table 4: Simplified Factorial Design

#	Factor	Level	Value
1	System type	1	IT enterprise
		2	Cyber-physical
2	Team experience level	2	>= 5 years
3	Time Allowed - Blue	1	2 weeks
4	Use of active blue defenders	2	Yes
5	Time Allowed - Red	1	1 week
6	Red Team Mimic	1	Mid-Tier
		2	High-Tier

11. Cost/Benefit Analysis

This section presents a simplified and informal analysis of the costs and benefits of developing, validating, and employing an analytic similar to the exemplar analytic discussed above. Costs are in US dollars with standard prefixes for K-thousand, M-million, and B-billion. We assume a single factorial approach, a target IT enterprise system, and a high-tier adversary. We begin by considering losses due to attacks. Then we tally the cost of running the experiment and estimate potential savings that such an analytic might bring.

Loss Estimates. Cybercrime loss estimates vary and can be difficult to quantify due to the many cost components (Figure 4). At the macro level, an illustrative example is the NotPetya attack of June 2017 [46] which affected government, energy, finance, defense, and other sectors across multiple countries, to include the United States. Global losses beyond those reported were estimated to be around $10B. The attack paralyzed the global shipping company Maersk, as the malware spread to terminal systems affecting their global shipping and logistics operations. The company estimated losses of between $200M and $300M [47]. Separately, Accenture [48] estimates a per organization annual cost in the $7M-$18M range.

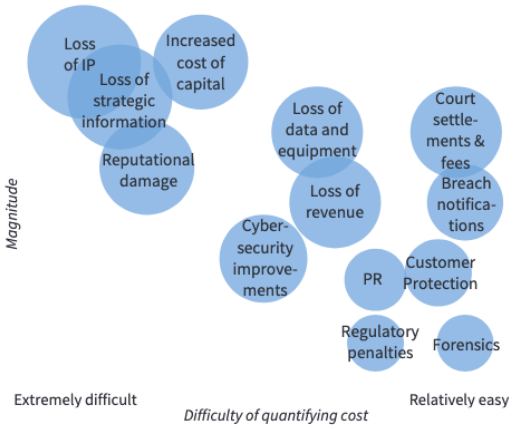

Figure 4: Cost Components of an Adverse Cyber Events (not exhaustive; e.g., loss of life not included) [49]

Analytic Development Costs. We use an analytic development cost of $5M, a figure we extrapolated from our own experience. The figure includes research and development of the analytic, as well as the development of related software for management of reference datasets, the user interface, and reporting. Also included are testing, documentation, and certain overhead costs (e.g., status reporting/oversight).

Analytic Validation Costs. The cost of conducting a single run of the treatment and control lines for the experimental design in Figure 3 is on the order of $220K, with labor details given in Table 5.

Table 5: Time to Run Experiment One Time

		Treatment	Control
Blue Team-E	Members	2	2
	Weeks	5	4
Red Team	Members	2	2
	Weeks	2	2
Blue Team-D	Members	2	2
	Weeks	1	1
User Team	Members	2	2
	Weeks	1	1
Total hours		**720**	**640**
Coordination hours		160	
Grand total hours		**1520**	

The table assumes a 40-hour work week and an average fully burdened hourly rate of $150.00. The human intensity of cyber is apparent, with four separate teams involved. The engineering blue team (Blue Team-E) has two members and is given four weeks to engineer protections into a target system. The treatment team is given an additional week of training on the analytic. The experiment itself takes place over one week.

The red team has two members and has two weeks to work, one for reconnaissance and one "on target." The defending blue team (Blue Team-D) has two members and actively defends the system during the experiment. For added realism, two users use the system during the test (and they may be a red team target).

If we run the validation experiment 30 times to achieve a minimum level of statistical power, the experiment cost is on the order of $6.6M. When added to the analytic development cost, the total is $11.6M. However, if we also assume that only one in four developed analytics (the exemplar analytic being just one tool) will prove to be effective, then we spend in excess of $46.4M to arrive at a single validated analytic. We further assume that the "shelf life" of the validated analytic is five years. Shelf life can be limited because threat actors may change their behavior over time sufficiently to reduce/negate the effectiveness of the analytic and accuracy of associated reference datasets.

Simplified Cost/Benefit Model. If, per the Accenture estimates [48], we assume that the average large organization suffers an annual loss of $15M and we further conservatively assume that (1) application of the analytic can reduce those costs by 5% yielding annual saving per organization of $750K and (2) the analytic costs $50K for annual licensing/training/support, then the annual savings accruing from use of the analytic is $700K or $3.5M over five years for a single organization. If 300 organizations employ the analytic, then the aggregate net five-year reduction in losses is roughly $1B. For 300 organizations at $50K per organization, revenue from the analytic is $75M. Table 6 summarizes these figures. The model suggests a non-trivial overall benefit based on the net payoff.

Table 6: Net Loss Reduction / Benefit

Single Organization	
Annual loss to cyber crime	15,000,000
Percent reduction of loss via analytic use	5%
Annual losses avoided	750,000
Annual analytic licensing/support cost	50,000
Annual net savings	700,000
5 year net savings	3,500,000
Multiple organizations	
Number organizations using analytic	300
5 year savings overall	1,050,000,000
Less development/validation cost	46,400,000
Net loss reduction / benefit	**1,003,600,000**

While we made several conservative assumptions (e.g., only one in four analytics successfully validate, yield only 5% savings, analytic shelf life is limited to five years, and only 300 organizations worldwide employ the validated analytic), we were not as conservative in other areas. Specifically, we assumed team sizes of two people each for blue/red/user teams. Teams are often larger, so doubling team sizes would nearly double the overall validation cost (a research question for future work is what an "optimal" team size actually is and what variables affect that size). In addition, we only allowed red teams one week to understand and surveil the target system. Sophisticated offensive cyber actors "in the wild" may spend months or years. However, in recognition of this fact, red teams are often given advance target information to help offset limited research time [6].

Three additional factors merit consideration. First, we did not address *how much* of an improvement (decrease) in CA is required in order to call the analytic a success. We merely stated the inequality and remained tacit on the magnitude, also an area of possible future work. Second, to strengthen external validity, one may wish to increase the factorial combinations tested (e.g., to consider different types of cyber systems). Lastly, there is the question of the labor pool from which to draw the teams in the experiment, particularly the red team labor pool. Recent experience has taught us that red team members tend to be in constant demand. Thus, the timely availability of skilled teams for participation in experiments of the type we described in this paper suggests the need for a "red team reserve force." The costs of assembling and maintaining such a cadre is a factor, though we note that such costs can be amortized across many validation experiments. Another possible way to reduce experimental costs and cybersecurity costs more generally is the use of automated red teaming algorithms that might serve as a proxy for human red teams. Work is on-going in this space (e.g., [50][51]) and the results may eventually become suitable for mimicking lower tier threats.

12. Conclusion

This paper described an approach for rigorously validating a cybersecurity analytic and the costs/benefits of doing so. As discussed, clear challenges to experimentation in cyber exist, including grappling with an immense state space [28], controlling for confounds, and affording experimentation costs.

Much of the variability that must be controlled relates to the human/social dimension of cybersecurity. However, such challenges are not unique to cyber. The social sciences have found ways to deal with such variability. As to the investment required, Schneider justified investment in a robust science of cybersecurity this way, "We know what to do when somebody

breaks a finger, and each year we create a new influenza vaccine in anticipation of the flu season to come. But only after making significant investments in basic medical sciences are we starting to understand the mechanisms by which cancers grow, and a cure seems to require that kind of deep understanding" [24]. We argue that the cyber threat is more like cancer than a broken finger. Tichy [29] argued for increased formal experimentation in computer science some 23 years ago. Many would argue that cybersecurity needs to heed that advice to advance the field at a fundamental level. Some may argue that the future human footprint in cyber may drop as a function of increased use of AI-driven "bots" that take on various cyber roles. While this is certainly possible, the bots themselves must be specified and trained by humans; therefore, we believe that the heavy human dimension to cyber will persist well into the future.

Alas, our team lacked the funding to execute the experiment described. As discussed earlier, validation at this level faces the cultural headwinds that have tended to prioritize short-term action over long-term scientific results, even if that action can lead to a false sense of security. To mitigate those headwinds, we advocate for a focused partnership between government, academia, and industry to help jump-start a tradition of strong validation in cybersecurity with hopes of spurring more interest from all involved parties given the potential payoffs.

13. References

[1] R. A. Maxion, T. A. Longstaff, and J. McHugh, "Why is There No Science in Cyber Science?: A Panel Discussion at NSPW 2010," in *Proceedings of the 2010 Workshop on New Security Paradigms*, 2010, pp. 1–6.

[2] C. Herley and P. C. van Oorschot, "Science of Security: Combining Theory and Measurement to Reflect the Observable," *IEEE Secur. Priv.*, vol. 16, no. 1, 2018.

[3] National Institute of Standards and Technology, "National Institute of Standards and Technology 800-30: Guide for Conducting Risk Assessments," 2012.

[4] T. Llanso, G. Tally, M. Silberglitt, and T. Anderson, "Applicability Of Mission-Based Analysis For Assessing Cyber Risk In Critical Infrastructure Systems," in *International Federation for Information Processing (IFIP) - Critical Infrastructure Protection VII*, 2013th ed., vol. VII, Springer Berlin Heidelberg New York, 2013, pp. 135–148.

[5] J. Hallberg, J. Bengtsson, N. Hallberg, H. Karlzén, and T. Sommestad, "The Significance of Information Security Risk Assessments Exploring the Consensus of Raters' Perceptions of Probability and Severity," in *International Conference on Security and Management*, 2017.

[6] M. McNeil and T. Llansó, "An Analysis of Adversarial Cyber Testing Practice," in *IEEE System Security Symposium*, 2020, p. 8.

[7] T. Llansó, "A Capability-Centric Approach to Cyber Risk Assessment and Mitigation," Dakota State University, 2018.

[8] MiniTab, "Accuracy vs. Precision: What's the Difference?," *MiniTab Blog*, 2012. [Online]. Available: https://blog.minitab.com/en/real-world-quality-improvement/accuracy-vs-precision-whats-the-difference.

[9] S. E. Maxwell, M. Y. Lau, and G. S. Howard, "Is psychology suffering from a replication crisis? What does 'failure to replicate' really mean?," *American Psychologist*, vol. 70, no. 6. American Psychological Association, 2015.

[10] B. J. Wiggins and C. D. Chrisopherson, "The replication crisis in psychology: An overview for theoretical and philosophical psychology.," *J. Theor. Philos. Psychol.*, vol. 39, no. 4, 2019.

[11] T. Longstaff, D. Balenson, and M. Matties, "Barriers to science in security," *Proc. 26th Annu. Comput. Secur. Appl. Conf.*, 2010.

[12] ISO/IEC/IEEE, "ISO/IEC/IEEE 24765:2017 Systems and Software Engineering – Vocabulary." 2017.

[13] B. W. Boehm, "Guidelines for Verifying and Validating Software Requirements and Design Specifications," in *EURO IFIP*, 1979.

[14] W. Trochim and J. Donnelly, *The Research Methods Knowledge Base*, 3rd ed. Atomic Dog Publishing Inc, 2006.

[15] B. Y. C. Collberg, T. A. Proebsting, W. H. En, C. Collberg, and T. A. Proebsting, "Repeatability in Computer Systems," *Commun. ACM*, vol. 59, no. 3, 2016.

[16] J. Vitek and T. Kalibera, "Repeatability, reproducibility, and rigor in systems research," *Proc. ninth ACM Int. Conf. Embed. Softw. - EMSOFT '11*, p. 33, 2011.

[17] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Q.*, vol. 13, no. 3, pp. 319–340, 1989.

[18] V. Venkatesh and F. D. Davis, "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies," *Manage. Sci.*, vol. 46, no. 2, Feb. 2000.

[19] W. H. DeLone and E. R. McLean, "Information Systems Success : The Quest for the Dependent Variable," *Inf. Syst. Res.*, vol. 3, no. 1, pp. 60–95, 1992.

[20] W. H. DeLone and E. R. Mclean, "The DeLone and McLean Model of Information Success : Systems A Ten-Year Update," *J. Manag. Inf. Syst.*, vol. 19, no. 4, 2003.

[21] C. Landwehr, "Cybersecurity: From engineering to science," *Int. Conf. Eng. Reconfigurable Syst. Algorithms*, 2011.

[22] D. McMorrow, "Science of Cybersecurity," *Sci. Cybersecurity a Roadmap to Res.*, vol. 7508, no.

November, pp. 1–45, 2010.
[23] T. Longstaff, "Barriers to Achieving a Science of Cybersecurity," *Next Wave*, vol. 19, no. 4, pp. 14–15, 2012.
[24] F. B. Schneider, "Blueprint for a Science of Cybersecurity," *Next Wave*, vol. 19, no. 2, pp. 47–57, 2012.
[25] A. Kott, "Towards Fundamental Science of Cyber Security," in *Network Science in Cybersecurity*, Springer, 2014.
[26] A. Kott, "Science of Cyber Security as a System of Models and Problems," in *Network Science and Cybersecurity*, 2015.
[27] S. Peisert and M. Bishop, "How to Design Computer Security Experiments," in *Fifth World Conference on Information Security Education*, 2007, pp. 141–148.
[28] T. E. Carroll, T. Edgar, D. Manz, and F. L. Greitzer, "Realizing scientific methods for cyber security," *ACM Int. Conf. Proceeding Ser.*, pp. 19–24, 2012.
[29] W. F. Tichy, "Should computer scientists experiment more?," *Computer (Long. Beach. Calif).*, vol. 31, no. 5, pp. 32–40, 1998.
[30] K. P. L. Coopamootoo and T. Groß, "Cyber security and privacy experiments: A design and reporting toolkit," in *IFIP Advances in Information and Communication Technology*, vol. 526, Springer International Publishing, 2018, pp. 243–262.
[31] R. A. Maxion, T. A. Longstaff, and J. McHugh, "Why is there no science in cyber science?," *Proc. 2010 Work. New Secur. Paradig. - NSPW '10*, p. 1, 2010.
[32] D. L. Kewley and J. F. Bouchard, "DARPA information assurance program dynamic defense experiment summary," *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans.*, vol. 31, no. 4, 2001.
[33] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer, "The Emperor's New Security Indicators," in *Proceedings of the 2007 IEEE Symposium on Security and Privacy*, 2007.
[34] K. Ferguson-Walter *et al.*, "The Tularosa Study: An Experimental Design and Implementation to Quantify the Effectiveness of Cyber Deception," in *HICSS*, 2019.
[35] M. Weir, S. Aggarwal, M. Collins, and H. Stern, "Testing Metrics for Password Creation Policies by Attacking Large Sets of Revealed Passwords," in *Proceedings of the 17th ACM Conference on Computer and Communications Security (CCS)*, 2010.
[36] L. I. Millett, B. Fischhoff, and P. J. Weinberger, *Foundational cybersecurity research: Improving science, engineering, and institutions*. Washington, DC: National Academies Press, 2017.
[37] L. Tam, M. Glassman, and M. Vandenwauver, "The psychology of password management: a tradeoff between security and convenience," *Behav. {&} Inf. Technol.*, vol. 29, no. 3, 2010.
[38] Q. Hu, Z. Xu, T. Dinev, and H. Ling, "Does Deterrence Work in Reducing Information Security Policy Abuse by Employees?," *Commun. ACM*, vol. 54, no. 6, pp. 54–60, Jun. 2011.
[39] A. Vance, D. Eargle, K. Ouimet, and D. Straub, "Enhancing Password Security through Interactive Fear Appeals: A Web-Based Field Experiment," in *46th Hawaii International Conference on System Sciences (HICSS)*, 2013.
[40] M. McNeil, T. Llanso, and D. Pearson, "Application of Capability-Based Cyber Risk Assessment Methodology to a Space System," in *Hot Topics in the Science of Security Symposium*, 2018.
[41] H. Simon, *The sciences of the artificial, (third edition)*, 3rd ed., vol. 33, no. 5. Massachusetts Institute of Technology, 1997.
[42] K. Peffers, M. Rothenberger, T. Tuunanen, and R. Vaezi, "Design Science Research Evaluation," *Design Science Research in Information Systems. Advances in Theory and Practice*. Springer Berlin Heidelberg, 2012.
[43] A. R. Hevner and S. Chatterjee, *Design Research in Information Systems*. 2010.
[44] J. Venable, J. Pries-Heje, and R. Baskerville, "A Comprehensive Framework for Evaluation in Design Science Research," in *DESRIST*, 2012.
[45] J. Gosler and L. Von Thaer, "Resilient Military Systems and the Advanced Cyber Threat," 2013.
[46] A. Greenberg, "The Untold Story of NotPetya, the Most Devastating Cyberattack in History," *Wired*, 2018. [Online]. Available: www.wired.com.
[47] N. Lord, "The Cost of Malware Infection? For Maersk, $300 Million," *Digital Guardian*, 2020. [Online]. Available: https://digitalguardian.com/blog/cost-malware-infection-maersk-300-million.
[48] Accenture, "Ninth Annual Cost of Cybercrime Study," 2019.
[49] U. S. Government, "Economic Report of the President," 2018.
[50] D. Malzahn, Z. Birnbaum, and C. Wright-Hamor, "Automated Vulnerability Testing via Executable Attack Graphs," in *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, 2020, pp. 1–10.
[51] J. Plot, *Red Team in a Box (RTIB): Developing Automated Tools to Identify, Assess, And Expose Cybersecurity Vulnerabilities in Department of the Navy Systems*, no. September. Naval Postgraduate School, 2019.